Demystifying the PCIe® Plug-Unplug

Alex Umansky

Chief Architect, PCIe Solutions

IT PL, Huawei

Disclaimer

<u>Presentation Disclaimer</u>: All opinions, judgments, recommendations, etc. that are presented herein are the opinions of the presenter of the material and do not necessarily reflect the opinions of the PCI-SIG®.

Agenda

- **Plug/Unplug Definitions**
- **Up-to-date System Topologies**
- **System Level Challenges on Plug/Unplug**
- **Conclusions and Call for Action**

Pluggable Topology – Model 200x

- **Where**  **PCIe® slot**
- **What**  **PCIe device (bridges/switches discussion on next slides)**
- **How**  **Well-mannered**
  - Standard hot plug controller and correspondent software flows

How Everything Started?

PCI Hot-Plug Specification

Revision	Revision History	Date
1.0	Original issue.	October 6, 1997
1.1	Add PCI 2.2, PCI power management, PCI-X, and SMBus.	June 20, 2001

PCI Standard Hot-Plug Controller and Subsystem Specification

REVISION	REVISION HISTORY	DATE
1.0	Initial release.	06/20/01

PCI Express® Base Specification

Revision	Revision History	DATE
1.0	Initial release.	07/22/02
1.0a	Incorporated Errata C1-C66 and E1-E4.17.	04/15/03
1.1	Incorporated approved Errata and ECNs.	03/28/05

- **1997: PCI Hot Plug™ introduced**
- **2001: Hot Plug Controller and software model defined**
- **2002: PCIe supports native Hot Plug, implement hot plug registers for monitor and control:**
  - Device capabilities
  - Slot capability/control/status

Hot Plug messages removed on spec 1.1

Elements of Hot Plug (History)

Element	Purpose	Required or Optional
Indicators	Shows the power and attention state of the slot	Required, except as noted in Section 2.3
Manually-operated Retention Latch (MRL)	Holds add-in cards in place	Required
MRL Sensor	Allows the SHPC and system software to detect the MRL being opened	Required, except as noted in Section 2.3
Electromechanical Interlock	Prevents removal of add-in cards while slot is powered	Optional
Attention Button	Allows user to request hot-plug operations	Optional
Software User Interface	Allows user to request hot-plug operations	Required
Slot Numbering	Provides visual identification of slots	Required

Element	Purpose
Indicators	Show the power and attention state of the slot
Manually-operated Retention Latch (MRL)	Holds adapter in place
MRL Sensor	Allows the Port and system software to detect the MRL being opened
Electromechanical Interlock	Prevents removal of adapter from slot
Attention Button	Allows user to request hot-plug operations
Software User Interface	Allows user to request hot-plug operations
Slot Numbering	Provides visual identification of slots
Power Controller	Software-controlled electronic component or components that control power to a slot or adapter and monitor that power for fault conditions
Out-of-band Presence Detect	Method of determining physical presence of an adapter in a slot that does not rely on the Physical Layer

- Legacy of 25 years
- Same elements as in 2001 HotPlug spec
- Power controller added in Spec 1.1
- In-band (by physical layer) and out-of-band (PRSNT of connector-specific pins) presence defined
- **DPC/eDPC ECN first time introduces surprise/unmanaged unplug**

SHPC Based Hot removal

- **Flow of Hot Removal:**
  - The user press Attention Button to start hot removal
    - Alternately starts hot removal via Application
  - HPC sends the request to the hot plug driver in OS
  - OS quiesces Device Driver
  - Non posted requests to be completed
  - Slot power down
  - LEDS indicate transitions and states
  - User waits for software to complete a sequence

Graceful removal:

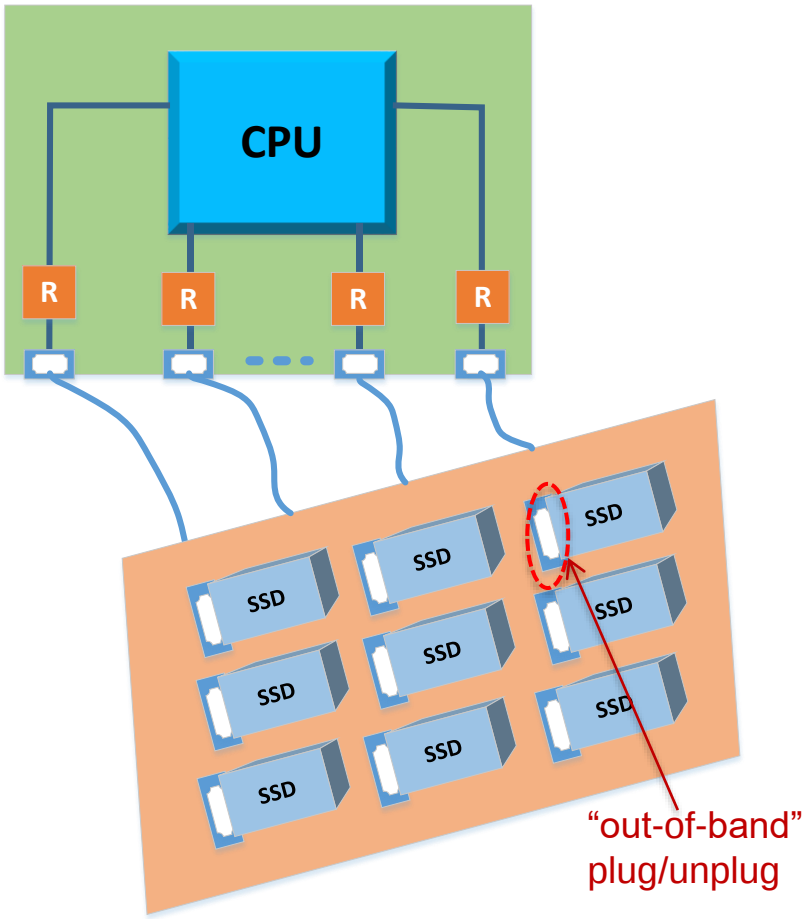
No transactions on link during hot unplug

SHPC – Cost

Slot Outputs	Slot Input
BUSEN#	PRSNT1#/PRSNT2#
RST#	PCIXCAP /M66EN
ATNLED#	MRL
PWRLED	BUTTON
CLKEN	FAULT#
System IO	
INT	SERR#/NMI#

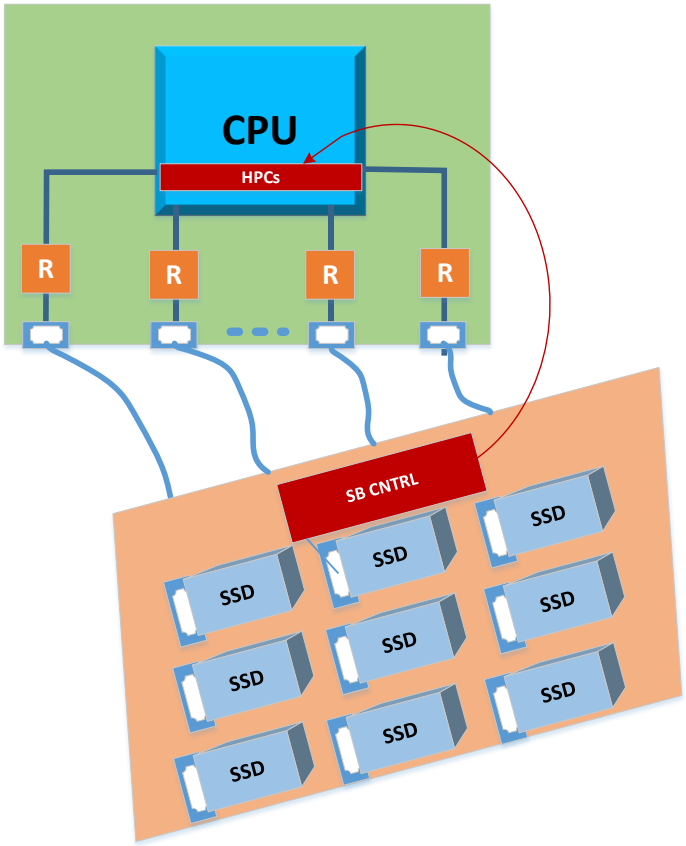
- >10 pins per port (part is optional, part is NA for PCIe)
  - 2017's system: 4 wires x2 Gen4 SSD

“Remote” Slot



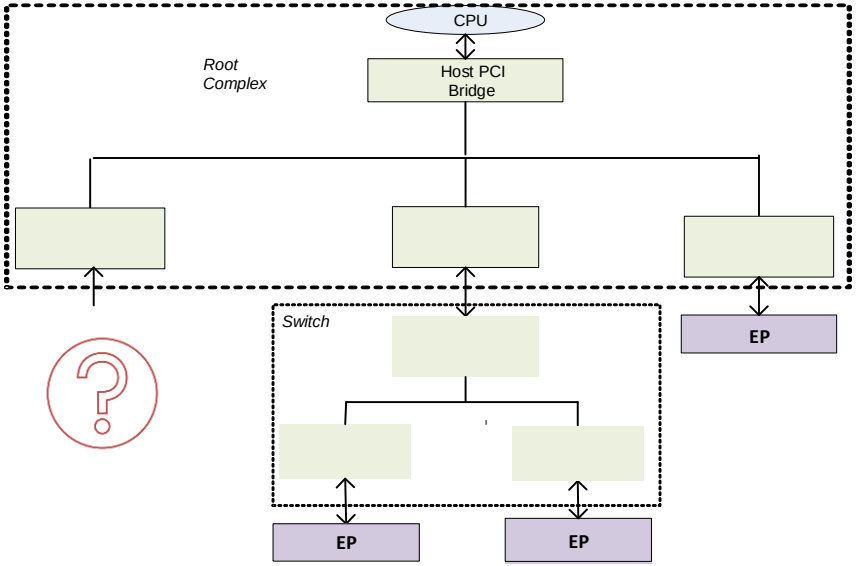
- **Device is not directly connected to Root or Switch Downstream port**
  - Example: storage devices on backplane
  - PRSNT# pins of port can’t “sense” the device
  - Physical link may travel thru re-timers/re-drivers
- **“In-band” presence can help, but**
  - L1 prevents in-band presence detection
    - Should be disable if in-band presence used
  - Unplug detection can be significantly slower than out-of-band

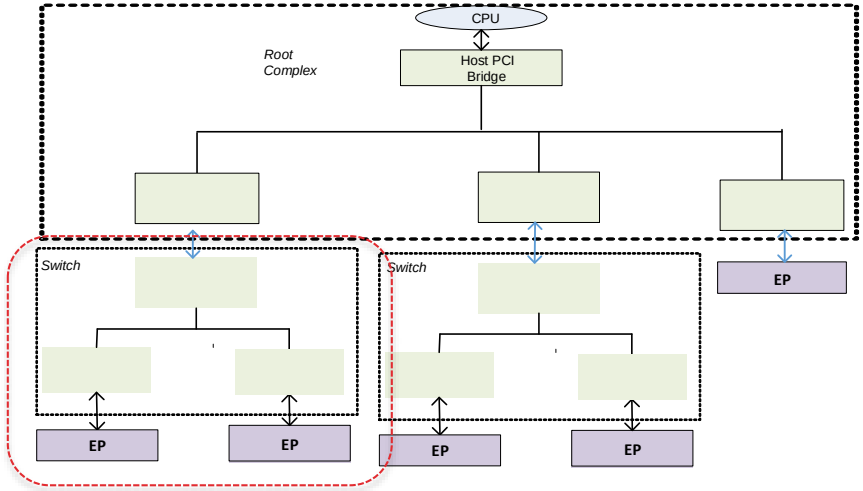
“Remote” Slot

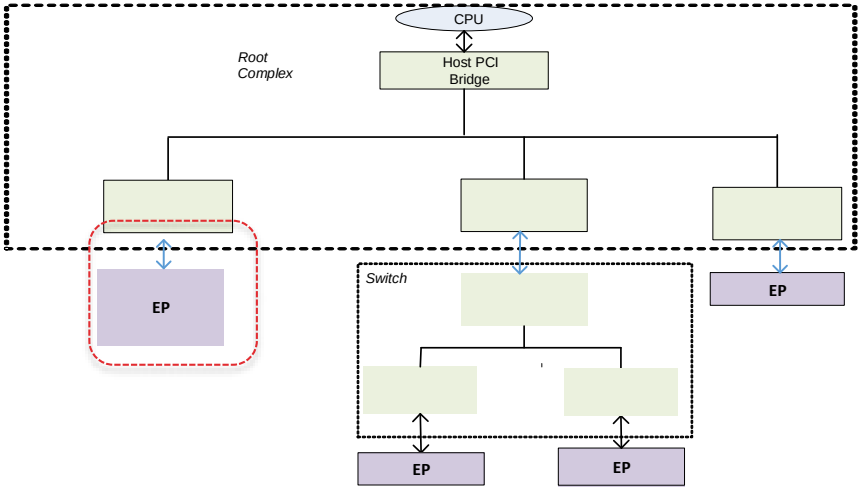


- Additional circuit on same board with “pluggable” devices monitoring status of slots (PRSNT#)
- On plug or unplug controls sideband HPCs in corresponded DP/RP
- Reaction time – longer that “direct” connection

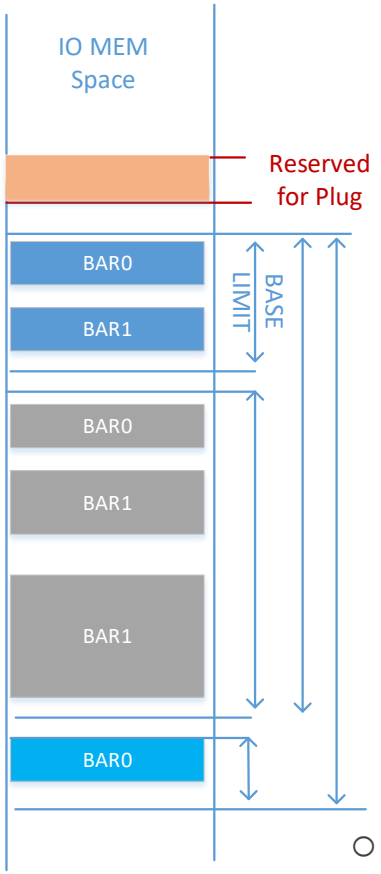
Hitless Hot Plug

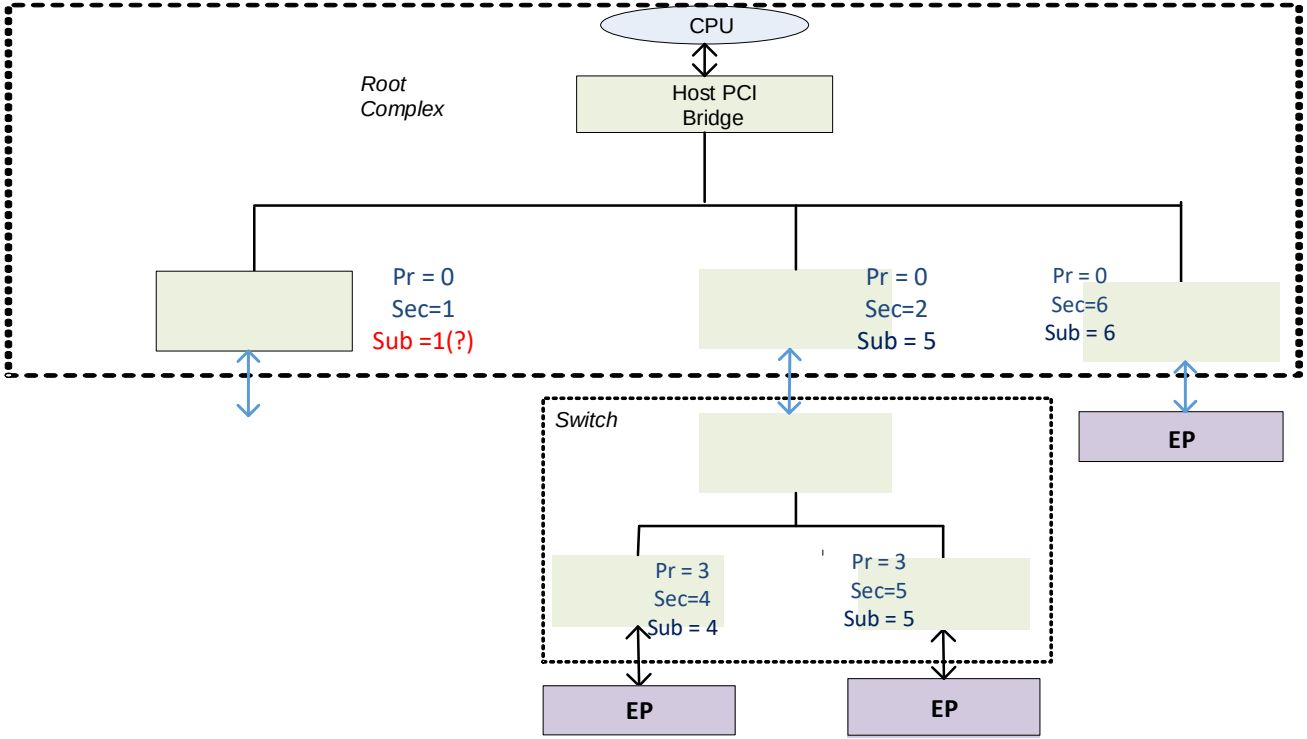






Resource Allocation



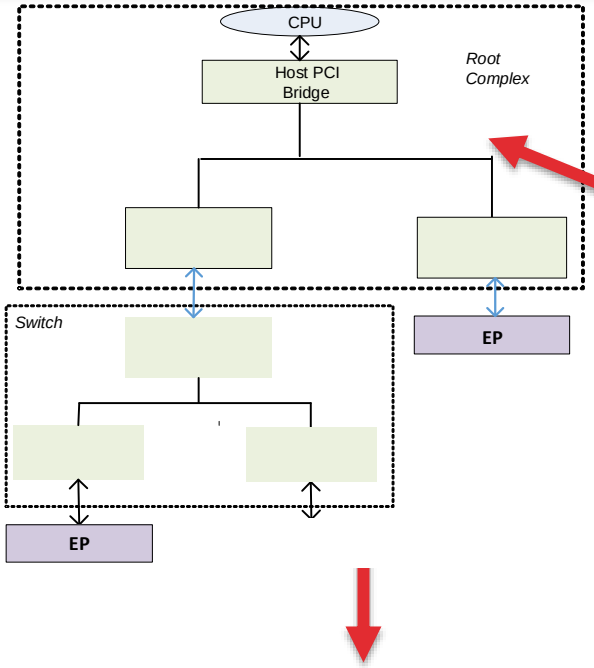


BIOS/EF should be tuned to reserve enough resources – bus numbers, mem space, interrupts

Unmanaged Unplug

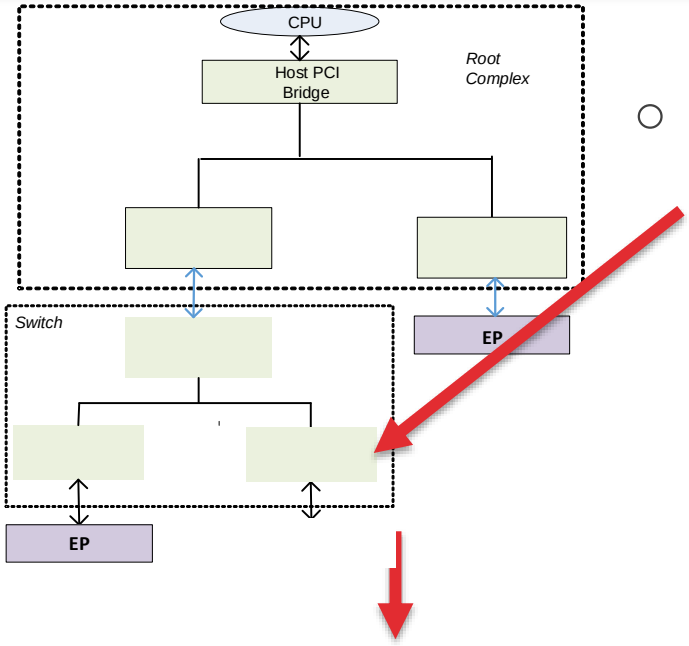
- **Unmanaged unplug – device removal without any prior notification to system**
- **System should be ready to manage this event without system crash and performance degradation for rest of the system (CPU/IO/memory)**

Unmanaged Unplug



- **RC error examples:**
  - Completion timeout – fatal
    - Blue screen before 4-5 years Modern server CPUs handle it
  - CPU pipeline crash – can't issue new commands due to stacked port

Unmanaged Unplug



- **DP and RP error examples:**
  - All layers may be affected:
    - Malformed packets
    - FC violations
    - TLP retransmissions
    - TLP loss
    - Symbol errors

Port's DPC should prevent error spreading and system crash

Removal Detection

- **In-band presence or sideband control of HPC in downstream ports are two methods to recognize unmanaged removal in modern storage/servers systems**
  - Tens → hundreds of devices/accelerators are located on backplanes and can be surprisingly removed
- **Removal recognition is important to prevent RP/Switches traffic blocking**

Removal Detection

- **Physical and Data Link errors may cause to transition of downstream port from L0 to Recovery state**
- **Link may stay in Recovery till the timeout – up to 48ms**
- **All this time Physical LINKUP, DL_LINK_UP and IN-BAND PRESENCE will be reported**

LTSSM State	Link Width	Link Speed	LinkUp	Link Training	Receiver Error	In-Band Presence[48]
Detect	Undefined	Undefined	0b	0b	No action	0b
Polling	Undefined	Set to 2.5 GT/s on entry from Detect. Link speed may change on entry to Polling.Compliance.	0b	0b	No action	1b
Configuration	Set	No action	0b/1b[49]	1b	Set on 8b/10b Error. Optional: Set on Link Error when using 128b/130b encoding.	1b
Recovery	No action	Set to new speed when speed changes	1b	1b	Optionally set on Link Error.	1b
~~L0~~	~~No action~~	~~No action~~	~~1b~~	~~0b~~	~~Set on Link Error~~	~~1b~~
L0s	No action	No action	1b	0b	No action	1b
L1	No action	No action	1b	0b	No action	1b
L2	No action	No action	1b	0b	No action	1b
Disabled	Undefined	Undefined	0b	0b	Optional: Set on 8b/10b Error	1b
Loopback	No action	Link speed may change on entry to Loopback from Configuration.	0b	0b	No action	1b
Hot Reset	No action	No action	0b	0b	Optional: Set on 8b/10b Error	1b

Blocking Prevent

- **Both RC and Switches should prevent buffer overflow during Removal detection to allow rest of IOs work without bandwidth drop**

Conclusions and Call for Actions

- **BIOS/EF/OS should be ready for plug of multi-hierarchy entity (device with embedded switch/ “lots of function” device occupying multiply buses)**
  - Hot-Plug Capable (Slot Capabilities) info may be not enough . New mechanism to allow Enumeration SW reserve Bus numbers / Memory space for potentially plug-in entities
- **Relaxation in device readiness time (second’s) required for new type of devices with functionality defined by internal SW = programmable engines / firmware**
- **To separate in-band and out-of-band presence detect enabling and reporting for different system topologies**
  - Sole In-band detection lets to handle a case of “logically” stopped device that continue to be physically attached to the slot
- **RC and switches should ensure stable traffic for “stay in place” IOs during plug/unplug events**
  - Surprise unplug case is especially interesting due to relative long “transition“ time via Recovery state

Thank you for attending the PCI-SIG Developers Conference Israel 2017.

For more information please go to www.pcisig.com